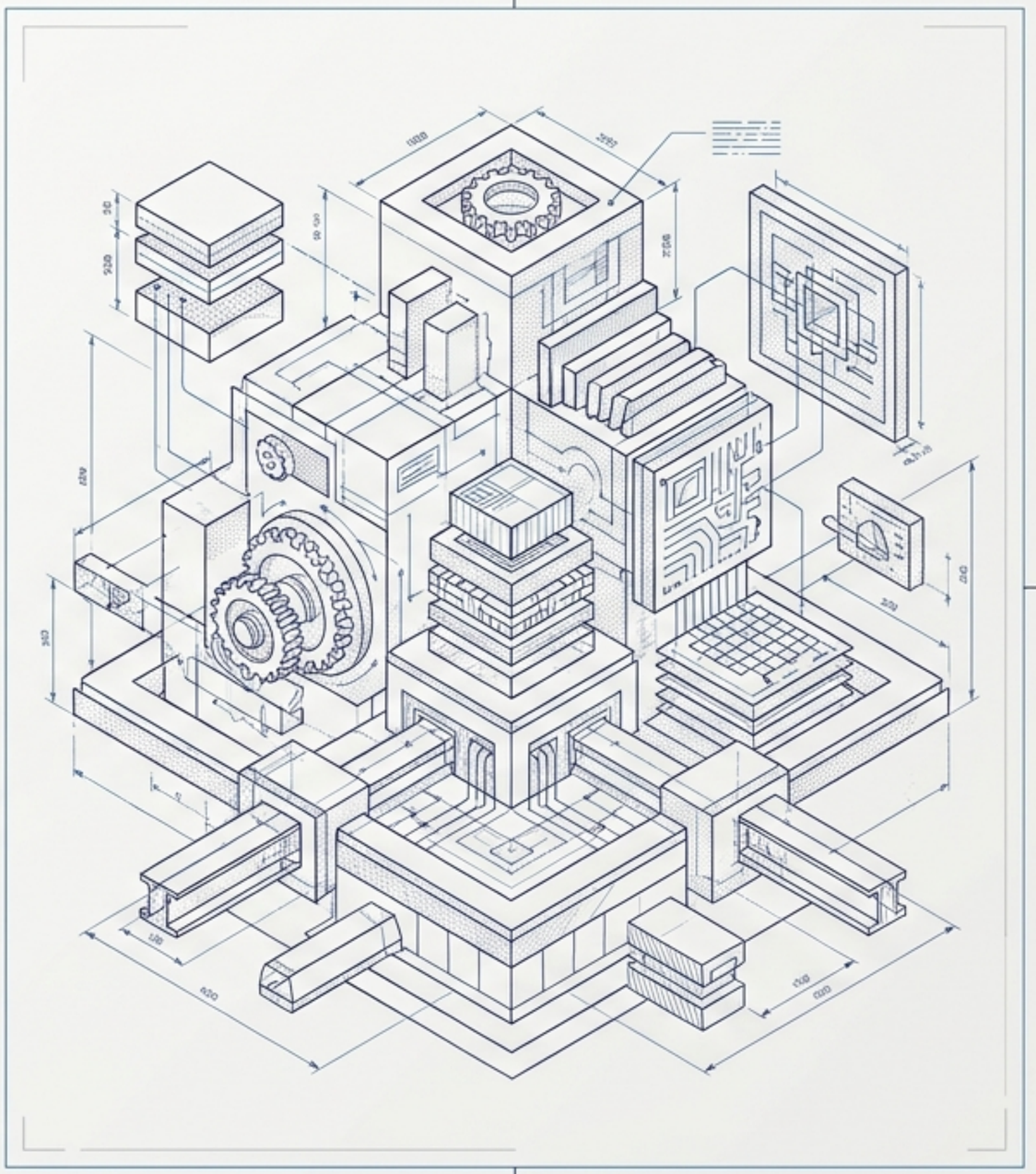
次世代モデル「MiniMax M3」の全貌

深層アーキテクチャの解剖と知財・特許実務のパラダイムシフト



M3を定義する3つの特権的能力

超長文コンテキスト

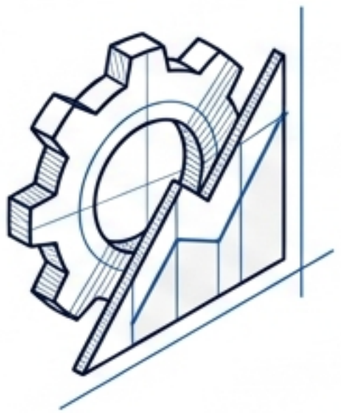
1,000,000 トークン

包袋全量や長尺動画を分割せずに一括処理。最低512Kを完全保証。

ネイティブ・マルチモーダル

テキスト・画像・動画の完全融合

視覚エンコーダの接合を排し、Step 0からの同時学習で深い意味解釈を実現。



フロンティア級推論

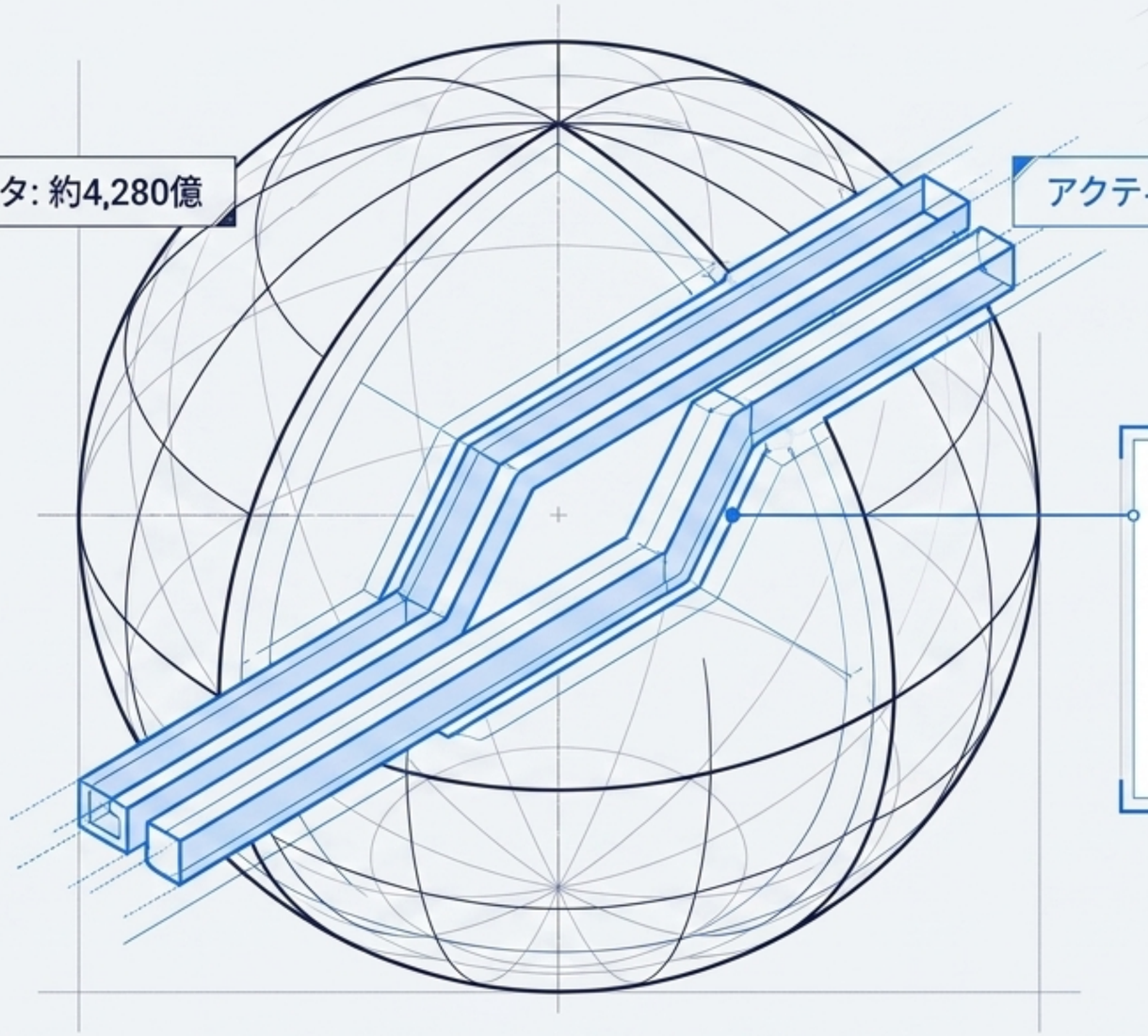
SWE-Bench Pro 59.0%

Gemini 3.1 Pro (54.2%) を凌駕し、最先端クローズドモデルに比肩する自律的実行力。

巨大空間と極限の効率：MoEアーキテクチャ

総パラメータ: 約4,280億

アクティブ: 約230億



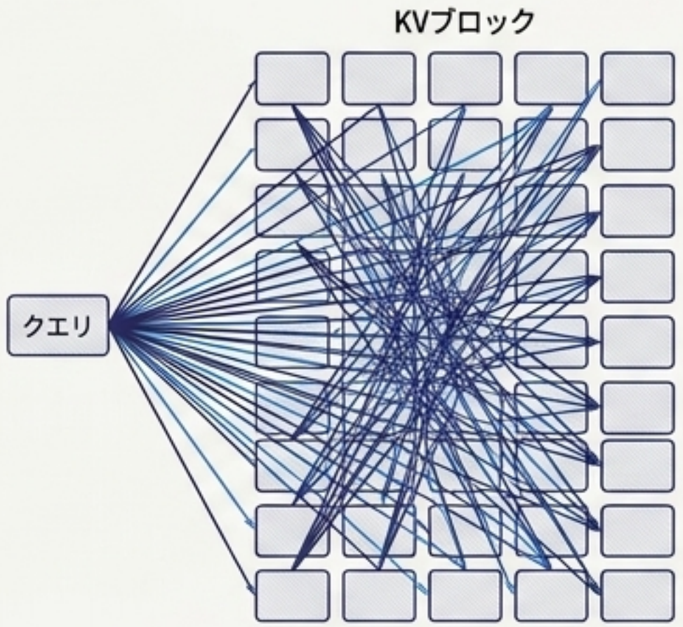
ネットワーク層数: 60層

データ精度: bfloat16

128個の独立エキスパート群から、1トークンあたり最も関連性の高い4つのみを動的にルーティング (4 active per token)。圧倒的知識量と極限の高速スループットを両立。

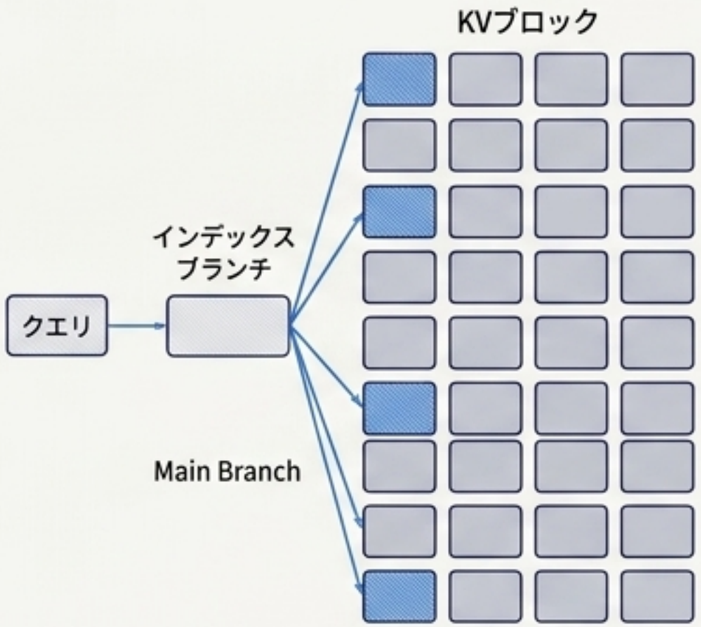
MSA：「二次関数的な壁」の突破

従来型アテンション



コンテキスト長の二乗に比例して計算量とメモリが爆発。

MSAアーキテクチャ（デュアルブランチ構造）



Index Branchで関連性の高いKVブロックを動的抽出し、Main Branchで選択された少数ブロックにのみ厳密な計算を実行。

FLOPs: 28倍削減	プレフィル速度: 9倍向上	⇒	デコード速度: 15倍向上

ネイティブマルチモーダル：視覚と論理の深層融合

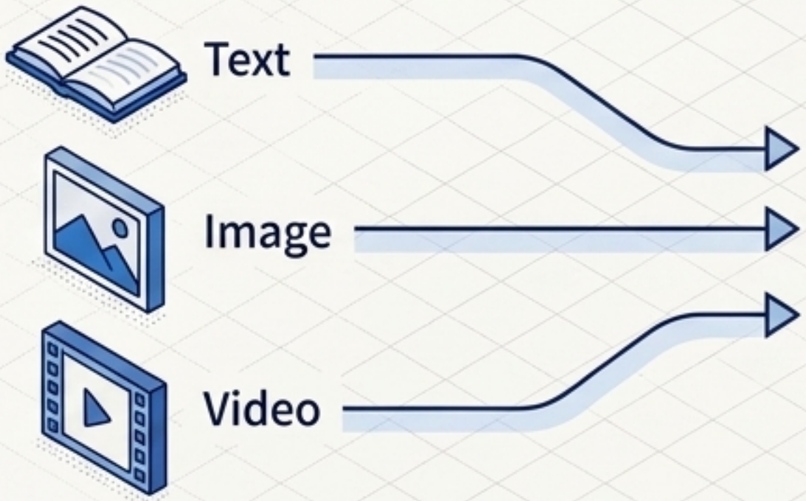
従来型 / 競合モデル

テキストのみで事前学習

ビジョンエンコーダを後付接合

画像とテキストの文脈にズレが生じ、ハルシネーション（幻覚）の温床に。

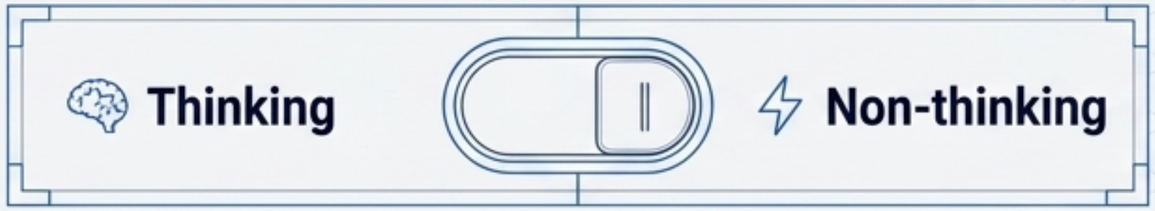
MiniMax M3



Step 0からのインターリーブ学習（同時訓練）

ピクセルの背後にある物理的構造・論理をテキストと同等の深さで理解。追加アダプタ不要で完全同期。

動的推論モード : "Thinking" vs "Non-thinking"



Thinking モード (思考)

メカニズム: 水面下で拡張された思考連鎖 (CoT) を展開

ユースケース: 複雑な特許クレーム解釈、長期的なFTO分析、マルチステップ・エージェント

ベネフィット: 限界まで引き出される論理推論能力

Non-thinking モード (非思考)

メカニズム: 最短経路で即時的な回答を生成

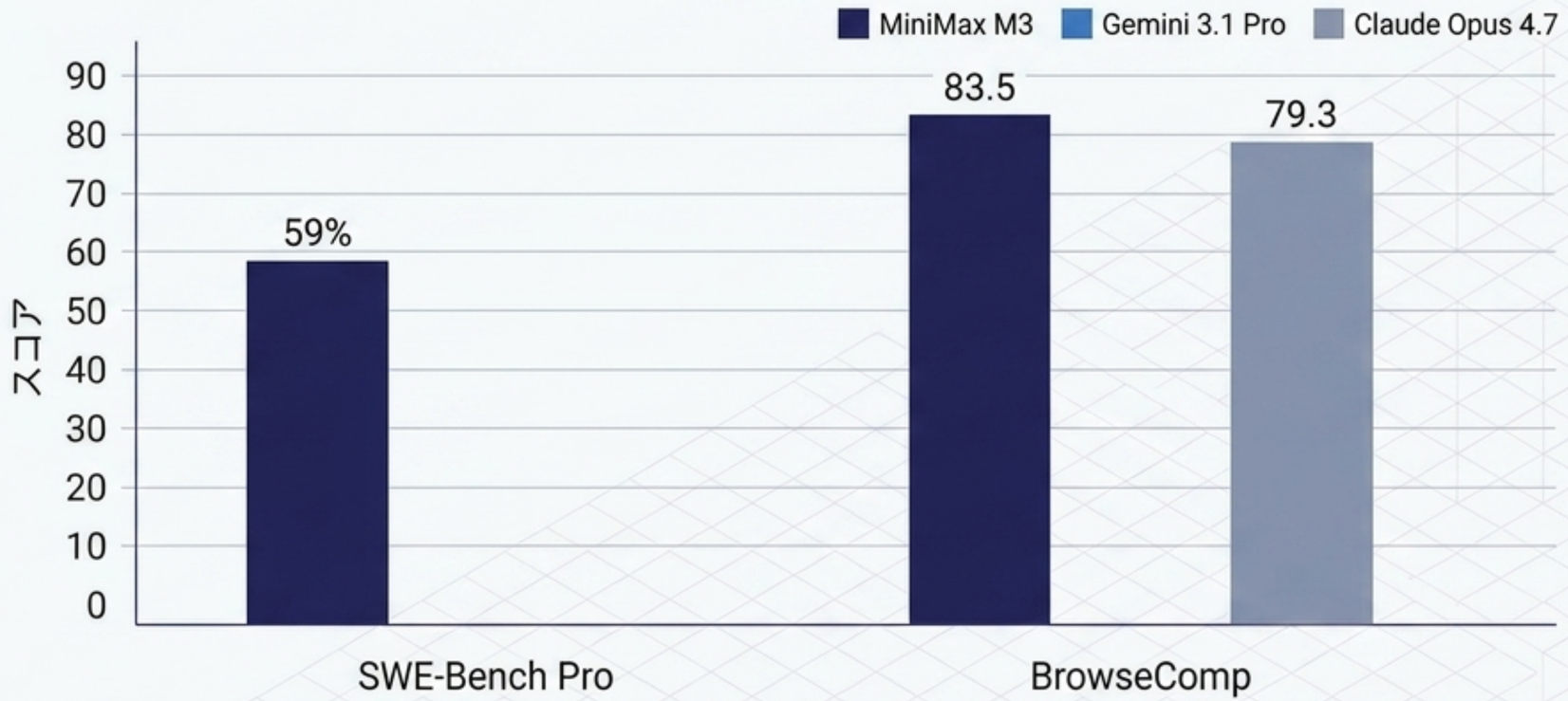
ユースケース: リアルタイムチャット、定型データ抽出、インラインコード補完

ベネフィット: 極限の低レイテンシ（応答速度）

タスクに応じてアーキテクチャの挙動を自由に切り替え可能。どちらのモードを使用してもトークン課金体系は同一。

圧倒的パフォーマンス：最先端クローズドモデルとの比較

主要ベンチマークにおけるMiniMax M3と最先端クローズドモデルの比較



単なる言語生成を超え、ソフトウェア開発の課題解決（SWE-Bench Pro）や自律的な情報探索（BrowseComp）において、現在市場の頂点にあるプロプライエタリモデルを凌駕、あるいは比肩する知能を証明。

自律型AIワーカーの証明：長期的安定性

Track 1: 研究代行 (12時間連続稼働)

タスク: ICLR 2025優秀論文の自律的再現

プロセス: 数式・チャート解析 → 1Mコンテキストへログ保持

成果: 介入ゼロで18回コミット、23図表出力（完璧な再現）

Track 2: システム最適化 (24時間連続稼働)

タスク: Hopper GPU FP8 GEMMカーネルの最適化

プロセス: 1,959回の自律ツール呼び出し → 147回のテスト

成果: ハードウェア使用率が7.6%から71.3%へ。実行速度9.4倍

破壊的コスト構造：APIエコノミクス

破壊的コスト構造：APIエコノミクス

モデル	入力コスト	出力コスト
MiniMax M3	$0.30	$1.20
Qwen 3.7 Plus	非公開	$1.60
Claude Opus 4.7	約10〜20倍	約10〜20倍

※100万トークンあたりの価格

The Game Changer

実効入力コスト: $0.101

OpenRouter統計でキャッシュヒット率平均83.1%。M3のキャッシュ読込コスト($0.06)により、特許包袋や巨大仕様書を反復処理する知財実務において、Opusの約5%〜10%のコストで運用可能。

知財業務における従来型AIの致命的ボトルネック

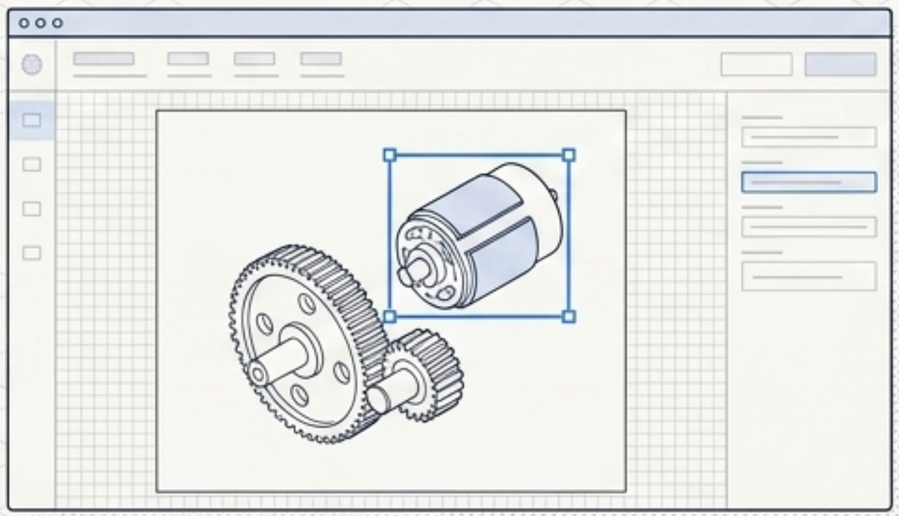
クレーム
（法的言語）

図面
（物理的・構造的表現）

特許文献において「クレーム」と「図面」は不可分である。従来のテキスト特化型LLMは図面の構造を想像するしかなく、参照符号の誤認や致命的なハルシネーションを発生させていた。

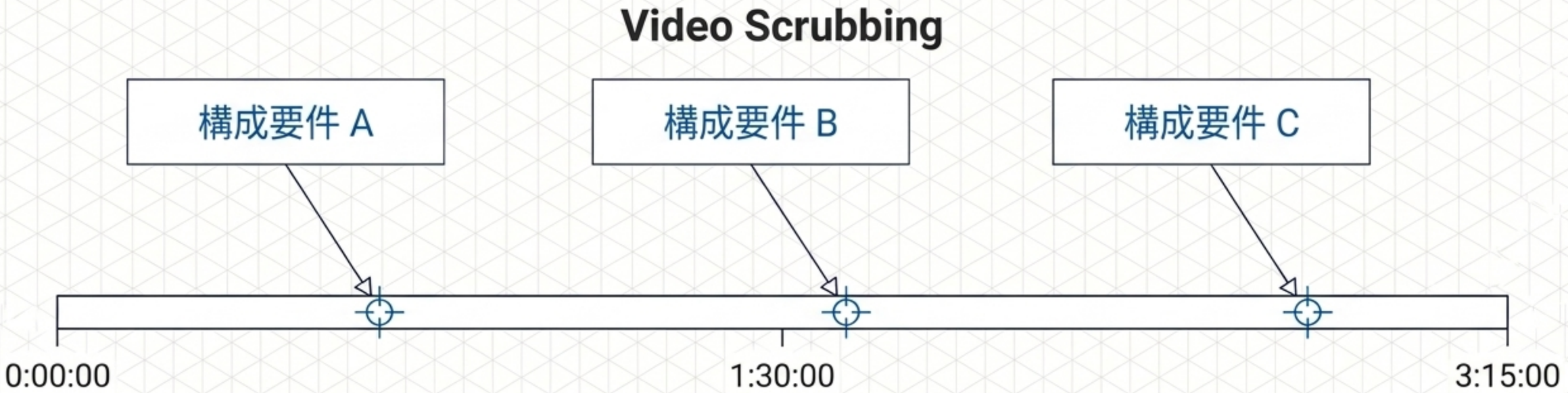
Solution: M3の「100万コンテキスト」×「ネイティブマルチモーダル」が情報分断の壁を完全に破壊する。

IP Revolution 1 - 明細書ドラフティングの自動化



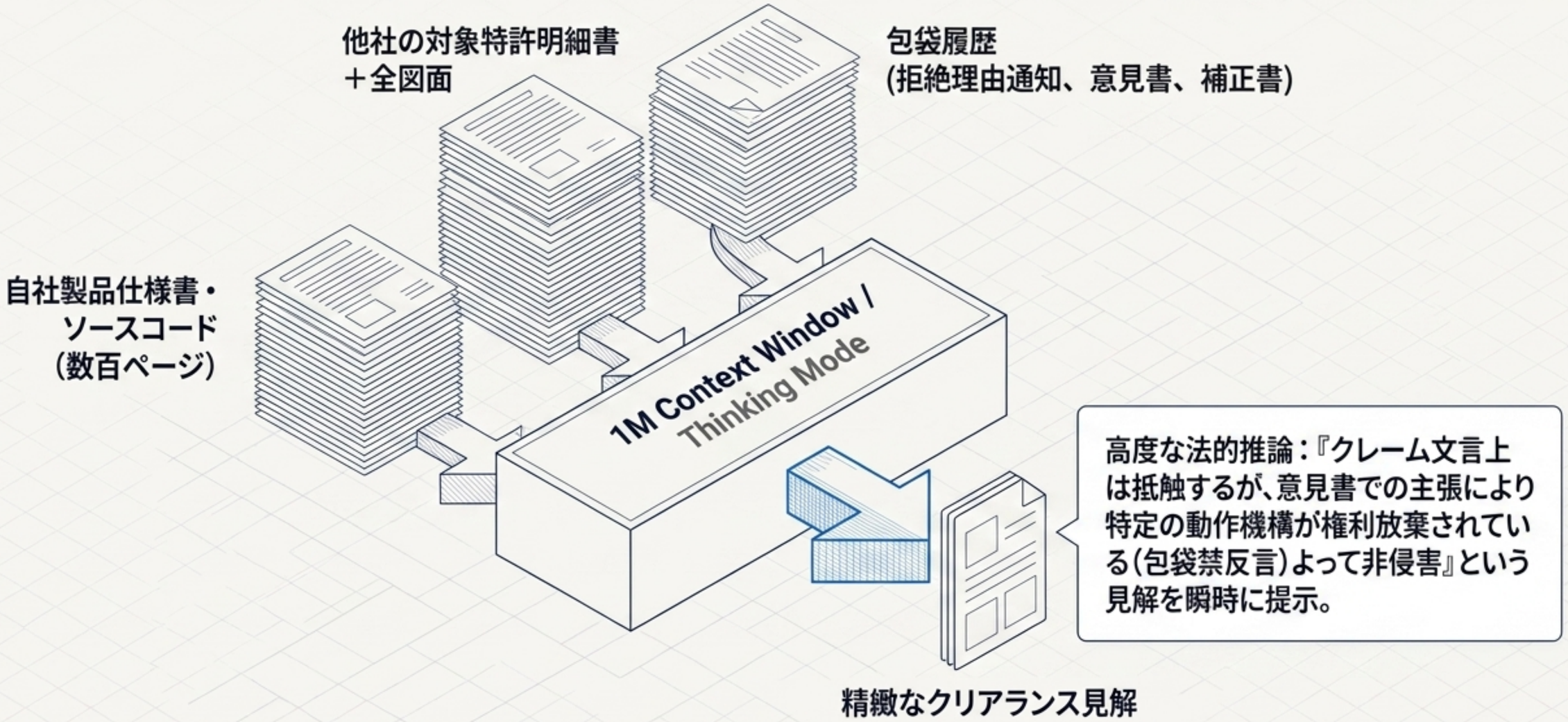
外部のOCRに依存せず、図面内の参照符号（例：「100: モーター」）とテキストの空間的関係性を直接理解。
画面上の部品をハイライトして従属クレームを追加する直感的な操作のバックエンドとして機能。

IP Revolution 2 - 動画先行技術（NPL）の全量調査



1. クレーム要件と数時間分の動画を同時にMコンテキストへ投入。
2. M3がピクセル変化と部品の動作順序を精査。
3. 全構成要件が揃う瞬間のタイムスタンプと物理的根拠を自律抽出。無効調査のパラダイムシフト。

IP Revolution 3 - FTOと包袋の全量解析

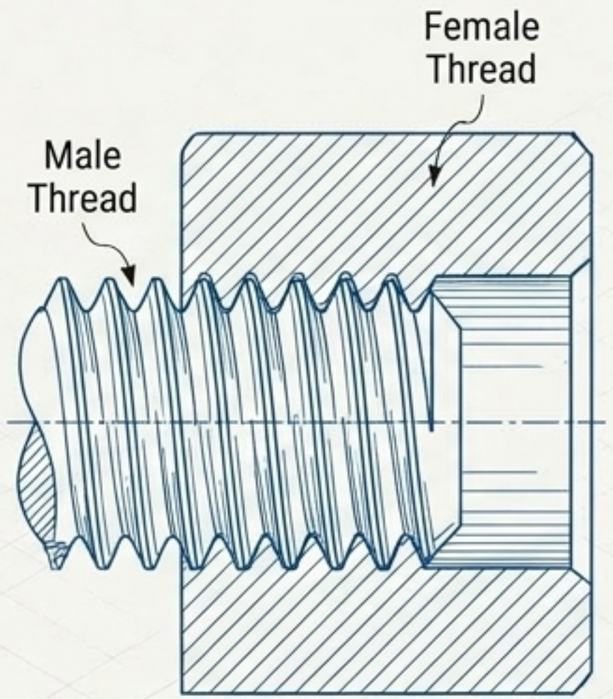


IP Revolution 4 - 文脈依存型マルチモーダル翻訳

従来型AI翻訳 - テキストのみ

...a connection part...

[Error] 単なる接続部と曖昧に翻訳され、致命的な権利範囲の縮減が発生。

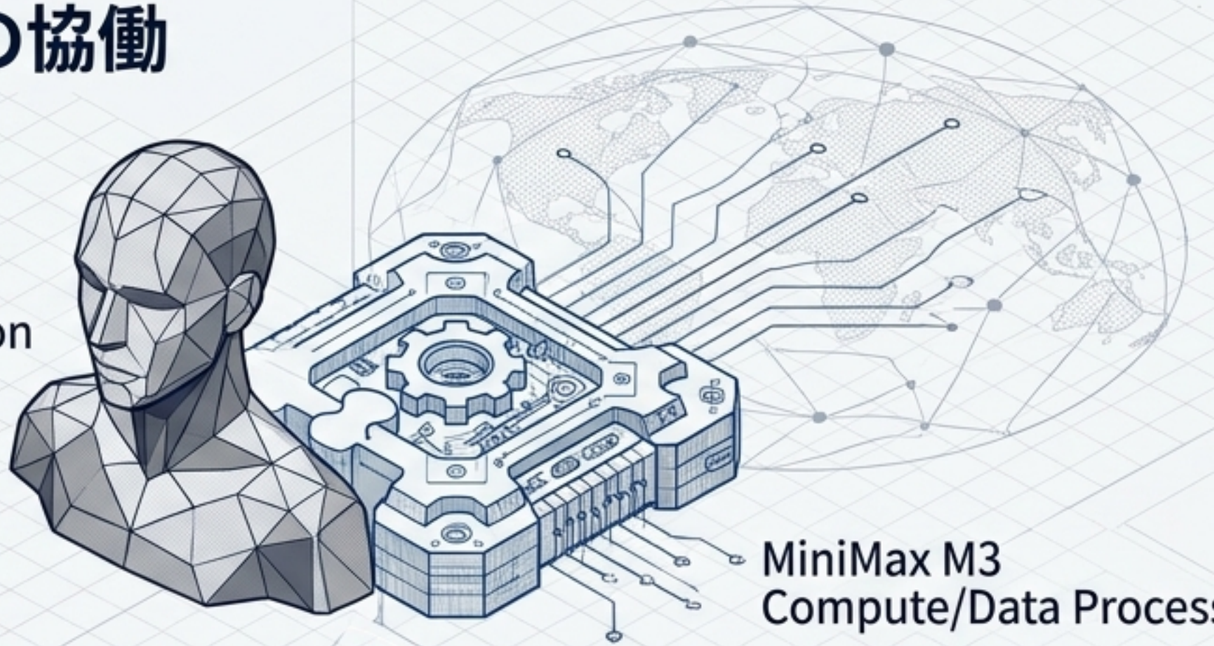


MiniMax M3 - 図面＋テキスト同時参照

...engagement of the male thread and the female thread...

[Success] 視覚情報から構造的文脈を補完し、技術的破綻のない翻訳を実行。グローバル出願の品質を担保。

未来の知財エコシステム：自律型AI知財ワーカーとの協働



MiniMax M3の役割	人間の専門家の役割
超高負荷な認知タスクの代行（包袋の全量解析、図面とクレームの同期推論、膨大なNPL動画の証拠発掘）。	より多角的で強固なグローバル知財戦略の立案、競合を無力化するクレーム網の設計、高度な法的折衝。

MiniMax M3は単なる効率化ツールではない。労働集約的なプロセスを消滅させ、知財部門の専門的リソースを「真の価値創造」へと集中させる次世代のインフラストラクチャである。